university of groningen

campus fryslân

Emotion Control in FastSpeech2-Based Speech Synthesis: Comparative Analysis of Prosody Scaling, Supervised Training, and Fine-Tuning

Hanyu Zhang

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Emotion Control in FastSpeech2-Based Speech Synthesis: Comparative Analysis of Prosody Scaling, Supervised Training, and Fine-Tuning

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Hanyu Zhang (S-5838975)

June 11, 2025

Acknowledgements

I would like to thank my supervisor Dr. Shekhar for his valuable guidance and support throughout this project. I am also grateful to Li Zhu for his helpful advice during my research.

I would like to acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster, which greatly facilitated the computational aspects of this research.

Finally, I sincerely thank my family and friends for their constant support and encouragement throughout this journey.

Abstract

Emotion is a critical component of human communication, and enabling synthetic speech to express emotion effectively remains a major challenge in modern text-to-speech (TTS) systems. This study investigates three modeling strategies within the FastSpeech2 framework: pitch and duration control, scratch training, and fine-tuning to assess their impact on the naturalness and emotional expressiveness of synthesized speech.

To evaluate these methods, three emotional categories (Sad, Angry, Happy) were synthesized using each modeling approach and assessed through a subjective listening test. The participantsants rated the naturalness of each sample on a five-point MOS scale and selected the most emotionally expressive version among the alternatives.

The results show that the fine-tuned model significantly outperforms the others, achieving the highest naturalness score (MOS = 4.47) and an emotion recognition accuracy of 72%. In contrast, the pitch-controlled and scratch-trained models scored lower and were not consistently perceived as emotionally expressive.

These findings demonstrate that fine-tuning with expressive data is the most effective and resource-efficient approach to building emotionally rich synthetic voices. The full demo and audio samples are publicly available at `https://burgundy07.github.io/emotion-demo/`

Keywords: Emotional speech synthesis, FastSpeech2, fine-tuning, MOS evaluation, prosody control

Contents

1 Introduction

Text-to-speech (TTS) synthesis, which converts written text into spoken language, has become a fundamental component of modern artificial intelligence systems.With advancements in deep neural architectures (especially sequence-to-sequence models), modern TTS systems have achieved significant improvements in naturalness and intelligibility. Prominent architectures,such as Tacotron and FastSpeech2 are capable of producing high-quality, human-like speech at scale (Ren et al., 2019; Wang et al., 2017). However, most conventional systems still generate speech with a neutral tone, lacking the emotional richness required for truly engaging and contextually appropriate human-computer interaction.

Emotional expressiveness in speech synthesis is essential for applications in socially interactive domains such as digital assistants, mental health tools, and storytelling systems.In these contexts, conveying empathy and intent through vocal tone enhances user trust and improves communication quality.While intelligibility and fluency are necessary, they alone are not sufficient.Achieving emotionally expressive speech requires variation in prosodic features such as pitch (F_0), duration, rhythm, and energy, which are challenging to model using standard TTS pipelines.

Recent research in emotional TTS focuses on two key challenges: achieving accurate emotional expression and enabling controllable emotional variation.Accuracy refers to whether the intended emotion (e.g., happiness, sadness, anger) is clearly perceived by listeners.Controllability, on the other hand, allows systems to vary emotional intensity, such as expressing mild sadness versus intense grief.For example, when synthesizing the sentence “I went out to see my friends today,” an emotional TTS system could deliver it joyfully or sadly, depending on context, and even modulate how strong the emotion sounds.

Two promising strategies have emerged to meet these challenges.First, prosody-based control adjusts acoustic parameters like pitch and duration at inference time to simulate specific emotions.This approach is interpretable and supports fine-grained manual control over prosody, making it suitable for dynamic emotional rendering (Lee & Kim, 2019). Second, embedding-based methods use low-dimensional emotion vectors extracted from reference speech to condition the synthesis model.These methods typically produce smoother prosody and more natural emotion expression, though they often sacrifice transparency and user control (Cornille, Wang, & Bekker, 2022).

However, existing research often evaluates these strategies in isolation, using different architectures, datasets, and evaluation protocols, making direct comparison difficult. To address this, the present study systematically compares pitch-based control, supervised emotion modeling (from scratch), and fine-tuning within the same FastSpeech2 framework, using consistent datasets and subjective and objective evaluation metrics.

Specifically, this research investigates how pitch and duration scaling, supervised emotion training, and transfer learning affect emotional clarity, naturalness, and efficiency across three emotional targets: happy, angry, and sad. The Emotional Speech Dataset (ESD) is used for emotional reference and training, while LJSpeech serves as the neutral baseline. The goal is to determine which method offers the best balance of interpretability, expressiveness, and training efficiency for emotional TTS

synthesis.
This paper is organized as follows. Section 1.1 introduces the research questions and hypotheses. Section 2 reviews relevant literature on emotional text-to-speech synthesis. Section 3 outlines the proposed methodology and implementation details. Section 4 describes the experimental setup. Section 5 presents and analyzes the evaluation results. Section 6 discusses key insights and their broader implications. Finally, Section 7 concludes the paper and suggests potential directions for future work.

1.1 Research Questions and Hypotheses

In light of prior work on emotional speech synthesis and the limitations of prosody control methods, this study addresses the following overarching research question:

> Can pitch-based prosody control and supervised emotion modeling (via from-scratch training and fine-tuning) enhance the emotional expressiveness of FastSpeech2-generated speech, and how do these methods differ in terms of naturalness, emotional clarity, and training efficiency?

This central question is decomposed into the following sub-questions:

- How does pitch and duration scaling compare to supervised emotion modeling (from-scratch or fine-tuning) in generating emotionally expressive speech?
- Which method—pitch-based control, from-scratch modeling, or fine-tuning—produces speech with more natural and emotionally fluent prosody?than pitch-based prosody manipulation?
- In terms of emotional clarity and data/resource usage, does fine-tuning a pre-trained FastSpeech2 model offer advantages over training emotion-specific models from scratch?

Based on these questions, the following hypotheses are proposed:

- H1: Both pitch-based prosody control and supervised emotion modeling (either via from-scratch training or fine-tuning) can enhance emotional expressiveness in synthesized speech.
- H2: Pitch scaling is expected to offer more transparent and interpretable control over emotional tone, while supervised emotion modeling is expected to produce more consistent and perceptually convincing emotional prosody.
- H3: Fine-tuning is expected to achieve emotional clarity comparable to or better than from-scratch models, with reduced training time and data requirements, indicating greater efficiency.

2 Literature Review

Speech synthesis has advanced significantly in recent years, especially with the emergence of neural models such as Tacotron and FastSpeech.Among these developments, emotional speech synthesis, which generates speech with expressive emotions such as happiness, sadness, and anger, has become a critical area of research due to its applications in human-computer interaction, virtual assistants, and affective computing.

To support this literature review, relevant papers were collected through targeted searches on Google Scholar, IEEE Xplore, ACL Anthology, and arXiv, focusing on works from 2000 to 2024. Keywords included "emotional speech synthesis," "prosody control," "emotion embeddings," "FastSpeech2,"and "phoneme alignment."Studies were included if they focused on neural TTS methods for emotional speech, presented original methods or evaluations, and reported subjective or objective results.Papers unrelated to emotion synthesis, lacking experiments, or presented as informal content were excluded.

This section reviews prior research in five key areas that underpin the emotional speech synthesis strategies explored in this study: (1) general approaches to emotional TTS, (2) prosody-based emotion control via pitch and duration modification, (3) emotion-specific training from scratch, (4) fine-tuning techniques for low-resource emotion transfer, and (5) methods for data alignment. These five themes directly inform the three experimental strategies compared in this work, namely prosody scaling, full model training, and fine-tuning, providing both conceptual foundations and practical benchmarks.

2.1 Background on Emotional TTS

Emotional text-to-speech (TTS) aims to synthesize speech that not only delivers linguistic content but also conveys the speaker's emotional state. This task differs from traditional TTS systems, which prioritize intelligibility and fluency, by requiring models to simulate affective prosody and expressive vocal features such as pitch contours, speaking rate, energy, and timbre. Emotional TTS has growing importance in human-computer interaction, conversational agents, and affective computing, where natural and emotionally appropriate responses improve user engagement and trust.

Early research on emotional speech synthesis was dominated by rule-based methods that applied manually crafted prosodic changes to neutral speech. For example, Schröder (2001) classified early approaches into rule-based, statistical, and transformation-based paradigms, noting that rule-based methods relied on linguistic heuristics (e.g., raising pitch and speeding tempo to express happiness) but were limited in scalability and naturalness. Statistical parametric speech synthesis (e.g., HMM-based TTS) introduced more data-driven modeling of prosodic features but often suffered from over-smoothing and lacked expressive variability (Burkhardt & Campbell, 2015).

A major shift occurred with the advent of deep learning and end-to-end architectures. Unlike modular pipelines, these architectures—such as Tacotron and FastSpeech—replaced hand-crafted modules with neural networks that learn alignment, duration, and acoustic features jointly. These models significantly improved synthesis quality, enabling more natural and data-driven expression of

emotional prosody. Kalita and Deb (2017)demonstrated that deep learning models outperform traditional methods in both naturalness and emotional clarity across languages and speakers. However, they also highlighted persistent limitations in speaker generalization and data efficiency, especially in multilingual or low-resource contexts.

Recent studies have increasingly highlighted the multidimensional nature of emotional expression. Traditional categorical approaches rely on discrete emotion labels , while dimensional models describe emotions along continuous axes such as valence and arousal. These models shape how emotional speech is represented and synthesized, influencing the choice of acoustic features and conditioning variables.

In addition, some studies emphasize the importance of subtle acoustic cues beyond pitch and duration. For instance, Hoult (2004) argues that emotional expression is closely related to spectral features such as formant movement, breathiness, and voice quality—elements that are often difficult to capture using conventional loss functions or coarse acoustic representations. This highlights the need for perceptually informed training objectives and more comprehensive models of expressive speech.

Taken together, emotional TTS systems have progressed from inflexible rule-based approaches to neural models capable of generating more natural and emotionally rich speech. Despite these advancements, key challenges remain, including limited data efficiency, difficulty in emotion control, and poor generalization across speakers and languages. To address these issues, this study explores three strategies—pitch-duration scaling, supervised emotion-specific training, and fine-tuning based on pre-trained models—within the FastSpeech2 framework, aiming to improve emotional expressiveness, controllability, and training efficiency.

2.2 Prosody Control Techniques

Prosody (which encompasses features such as pitch (F0), duration, and energy) is fundamental to expressing emotion in human speech. In emotional text-to-speech (TTS) systems, prosody control offers an interpretable and lightweight alternative to embedding-based or end-to-end methods, especially in low-resource emotional settings. Unlike deep emotion modeling approaches, prosody manipulation allows explicit modification of acoustic cues directly associated with affective states, such as raising pitch for excitement or lengthening duration for sadness.

A seminal model in this domain is FastPitch, which integrates a pitch predictor module into the synthesis pipeline, enabling direct F0 control during inference (Łańcucki, 2021). This architecture allows pitch contours to be globally adjusted without retraining the model. While this method is simple and does not require labeled emotional data, overly uniform adjustments may lack the nuanced dynamics found in natural emotional speech

Earlier rule-based approaches also explored pitch–duration coupling.Kim, soo Hahn, Yoo, and Bae (2008) proposed a method that controls pitch in the time domain and modifies duration in the frequency domain using PSOLA-based synthesis. They reported that the proposed algorithm obtained a higher MOS score for naturalness, achieving an average rating of 3.38. These findings suggest that

even rule-based, signal-level techniques can enhance perceptual expressiveness without requiring model retraining.

At a more granular level, Fahad, Singh, Gupta, Deepak, and Abhinav (2019)introduced a vowel-specific modification approach that dynamically adjusts F0, energy, and duration, rather than applying fixed global scaling. Their method showed up to a 13% improvement in subjective emotionality ratings, particularly for high-arousal emotions such as anger and fear.

Linguistic studies have shown that prosodic variation strongly correlates with emotional expression.Koike, Suzuki, and Saito (1998) , in a study on synthesized Japanese speech, found that joy and anger were characterized by higher pitch and shorter durations, whereas sorrow was associated with lower pitch and elongated syllables . These results support the intuition behind current pitch-scaling strategies in emotional speech synthesis.

From an implementation standpoint, phoneme-level control has been shown to be more effective than coarser, utterance-level prosody manipulation. Bulut et al. (2005) demonstrated that synchronizing prosodic modifications such as pitch and energy with phoneme boundaries, along with applying spectral envelope shaping, significantly enhanced emotional expressiveness and improved listener ratings.

Prosody modification plays a crucial role in expressive speech synthesis and voice conversion. Conventional methods, such as TD-PSOLA and fixed-factor epoch-based approaches, typically apply uniform scaling of pitch and duration across the entire utterance, limiting their ability to reflect the dynamic prosodic patterns characteristic of emotional speech. To overcome this limitation,Govind and Prasanna (2012) introduced a dynamic prosody modification method using zero-frequency filtered signals (ZFFS), which enables precise prosodic control at the level of individual epochs. This approach supports fine-grained, time-varying adjustments of pitch, duration, and excitation strength, providing greater flexibility and naturalness compared to traditional techniques.They demonstrated that this dynamic approach significantly outperformed fixed-factor methods in emotion conversion tasks, as supported by subjective evaluations: The subjective evaluations performed for the emotion conversion indicate the effectiveness of the dynamic prosody modification over the fixed prosody modification for emotion conversion.

In conclusion, prosody-based control offers a transparent and practical approach for emotional TTS. When applied at fine-grained units like phonemes or syllables, it not only enables better expressiveness in low-resource settings but also serves as a benchmark for evaluating more advanced emotion synthesis systems.

2.3 Emotion Modeling Approaches

Recent developments in emotional speech synthesis have moved beyond prosody-based adjustments, embracing data-driven techniques that condition generation on learned emotion representations. These emotion embeddings, typically extracted from reference audio, aim to capture holistic affective characteristics, such as pitch, rhythm, and energy, enabling models to produce speech that aligns

more closely with emotional intent. A widely used approach is Global Style Tokens (GST), initially designed for Tacotron, and later adapted into non-autoregressive frameworks like FastSpeech2.

Diatlova and Shutov (2023) extended the FastSpeech2 model by incorporating trainable emotion embeddings and a Conditional Cross-Attention (CCA) mechanism into both the encoder and decoder. This architecture enables token-level reweighting based on emotional context, allowing finer control over expressive variation in synthesized speech.The proposed model outperforms an existing implementation of FastSpeech2 extended for Emotional Speech Synthesis, regarding MOS and emotion recognition accuracy, without bringing inference speed latency

Fine-tuning, rather than training models from scratch, has proven especially valuable in low-resource emotional TTS. Kolekar, Richter, Bappi, and Kim (2024) demonstrated demonstrated that incorporating pitch, energy, and duration as emotion-related features into the variance adaptor of FastSpeech2 leads to high MOS performance (up to 4.09), especially when applied to multi-speaker and fine-tuned setups. Their method highlights the potential of fine-tuning pretrained models for scalable and expressive speech generation. This is echoed by Inoue, Zhou, Wang, and Li (2024), who introduced a hierarchical emotion distribution module in FastSpeech2 that allows fine-grained emotional control at the phoneme level, offering both high accuracy and interpretability in emotional TTS.

Nithin and Prakash (2022) conducted a direct comparison between Tacotron 2 and FastSpeech 2, both fine-tuned on emotional speech from the ESD corpus after pretraining on LJSpeech. Their results showed that, while Tacotron 2 achieved higher classification accuracy (90%) using the ScSer emotion recognition model, FastSpeech 2 converged faster and exhibited greater robustness to overfitting—especially in scenarios with limited training data per emotion. These findings support the use of transfer learning for emotional TTS, particularly when only a few hundred samples are available for each target category.

Together, these studies reinforce the effectiveness of embedding-based conditioning and fine-tuning as powerful, scalable approaches for generating expressive emotional speech. While less interpretable than rule-based prosody manipulation, these methods offer higher perceptual naturalness and are more robust under varying linguistic contexts.

2.4 Dataset and Alignment Techniques

High-quality data preprocessing and alignment are essential for the success of emotional text-to-speech (TTS) systems, especially in duration-sensitive architectures such as FastSpeech2. Unlike prosody control methods, which modify acoustic parameters at inference time, preprocessing aims to optimize input representations and maintain timing accuracy before model training.

One of the most foundational components of this stage is forced alignment, which ensures accurate mapping between audio signals and phoneme sequences. Tools like the Montreal Forced Aligner (MFA) are frequently employed to generate phoneme-level durations that guide TTS models during training. He, Sun, Zhu, and Zhao (2022) emphasized that alignment accuracy directly impacts synthesis quality, particularly when modeling emotional nuances that depend on subtle timing differences.

Emotion-specific data filtering is another essential preprocessing step. For instance, Cen, Dong, and Chan (2011) proposed emotion-aware data refinement by detecting and retaining only emotionally consistent utterances through automatic emotion recognition. This reduces noise in training corpora and ensures that models learn from expressive, emotionally salient examples.

Phoneme-level annotation is essential for alignment and training in emotional speech synthesis. As noted Tits, Haddad, and Dutoit (2019), The phonetic annotations are not time-aligned with our data yet, but methods can be used such as forced alignment systems. This suggests that phonetic annotations, even if not aligned, can still be used to support consistent training and potentially enable cross-corpus synthesis scenarios.

In terms of dataset construction, Thi, Thang Ta, Le, and Hai Do (2023) introduced an automated pipeline that combines pretrained models with publicly available corpora. Their method automates multiple stages in the construction of emotional speech datasets, significantly reducing manual effort and improving scalability, particularly in low-resource emotional categories and languages.

In conclusion, preprocessing steps such as alignment, phoneme normalization, and emotion-based data filtering are essential for building reliable and expressive emotional TTS systems. While these processes do not directly affect prosodic output, they lay the groundwork for accurate learning and robust generalization in emotional speech synthesis.

2.5 Summary and Challenges

Recent advances in emotional speech synthesis include prosody-controllable models and emotion embedding methods. However, these approaches have rarely been evaluated together in a unified experimental setting.

Łańcucki (2021) proposed FastPitch and similar models that add pitch predictors to allow real-time changes in F_0 during speech generation. This can improve emotional expression but, if not tuned well, may cause unnatural prosody. They warned that overly aggressive pitch control may reduce naturalness. Later models like PiCo-VITS demonstrate that using full pitch contours improves naturalness, especially for high-arousal emotions like anger, as shown by Wong and Chung (2024). In a similar vein, Inoue et al. (2024) proposed hierarchical pitch control mechanisms that operate at multiple linguistic levels, allowing for fine-grained prosodic variation.

Furthermore, advanced models such as MsEmoTTS enable multi-scale emotional transfer by combining sentence-level and word-level style conditioning. As stated by Lei, Yang, Wang, and Xie (2022), the proposed model is a unified and flexible model that allows us to synthesize emotional speech in different ways, including emotion transfer from reference audio, prediction from input text, and manual specification, thus offering broader expressive capabilities than GST-based systems.

Diatlova and Shutov (2023) extended FastSpeech2 with trainable emotion embeddings and a Conditional Cross-Attention mechanism, enabling token-level emotion reweighting. Their model outperformed baselines in MOS and classification accuracy, though evaluation reliability was limited by dataset subjectivity.

Despite recent advancements, achieving expressiveness in TTS remains a significant challenge. First, data dependency remains a significant limitation for embedding-based systems because these models often require high-quality emotional reference signals, which are difficult to obtain in low-resource or multilingual settings. Second, prosodic generalization continues to be a challenge for pitch-controlled methods, especially when dealing with linguistically diverse inputs that contain varying syntactic structures and prosodic patterns. Finally, the absence of standardized evaluation benchmarks makes it challenging to perform direct and systematic comparisons across emotional TTS systems with respect to emotional fidelity, naturalness, and controllability.

This study addresses these gaps by systematically evaluating pitch-control, scratch training, and fine-tuning-based emotional synthesis methods under a unified framework.A long-term research goal is to develop systems that integrate controllability, naturalness, and emotional expressiveness while maintaining generalization across speakers, languages, and domains.

3 Methodology

In this chapter, I describe the experimental setup employed for evaluating the emotional expressiveness of speech from FastSpeech2-based models. This entails the description of the model architecture, selection of the data and preprocessing, alignment methods, and methods for modeling emotions, as well as relevant ethical concerns.

3.1 FastSpeech2 Architecture

FastSpeech2 is a non-autoregressive text-to-speech (TTS) model that improves training efficiency and output quality compared to autoregressive architectures. It is composed of three primary modules: an encoder, a variance adaptor, and a decoder.

The encoder transforms phoneme sequences into hidden representations through stacked feed-forward Transformer blocks, incorporating multi-head self-attention and one-dimensional convolutional layers. These representations are passed to the variance adaptor, which integrates prosodic features—namely duration, pitch, and energy—into the sequence. The adaptor comprises three predictors: a duration predictor, a pitch predictor for fundamental frequency (F0), and an energy predictor for intensity variation. Each predictor includes two one-dimensional convolutional layers followed by ReLU activation, layer normalization, dropout, and a linear projection. Supervision is provided using ground-truth annotations of pitch, duration, and energy during training.

The decoder reconstructs mel-spectrograms from the prosody-enhanced representations in parallel. A neural vocoder such as HiFi-GAN is subsequently used to convert the mel-spectrograms into waveform audio Ren et al. (2022).

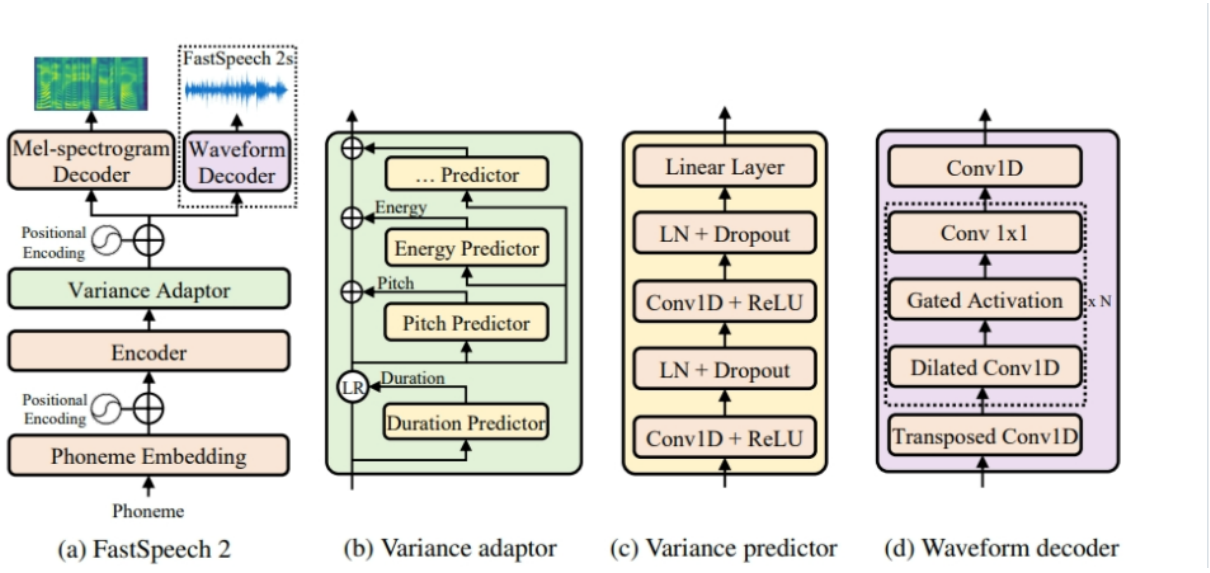

Figure 1: Core components of FastSpeech2. The model takes phoneme sequences as input and processes them through an encoder and variance adaptor that injects prosodic features (pitch, energy, duration). The output mel-spectrogram is converted to waveform using a vocoder.

3.2 Datasets

Two publicly available speech corpora were employed in this study:

- **LJSpeech:** A dataset consisting of 13,100 English utterances recorded by a single female speaker. This corpus was used for training the base FastSpeech2 model, representing neutral speech conditions.
- **Emotional Speech Dataset (ESD):** A multilingual corpus containing five emotion categories: Neutral, Happy, Sad, Angry, and Surprise. The English subset was selected for this study, and a single speaker (`spk_0015`) was used to ensure consistency. Approximately 350 utterances were sampled per emotion for Happy, Sad, and Angry Zhou, Chong, Wang, and Zeng (2022).

All audio samples were downsampled to 16 kHz, converted to mono, and normalized using peak normalization. Transcripts were processed through a grapheme-to-phoneme (G2P) converter to obtain phoneme sequences compatible with FastSpeech2.

3.3 Alignment and Duration Extraction

Accurate phoneme-to-frame alignment is essential for effective training of the duration predictor. The Montreal Forced Aligner (MFA) was used to generate phoneme-level alignments, producing Praat-compatible TextGrid files containing phoneme boundary annotations.(McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017).

Durations were extracted from these alignments and converted into frame-level units. These durations were subsequently used as training targets for the duration predictor, enabling accurate modeling of temporal prosody and speech rhythm.

3.4 Emotion Modeling Strategies

Three modeling strategies were explored to investigate their influence on emotional expressiveness in synthesized speech:

- **Pitch and Duration Control:** Emotional expressiveness is simulated during inference by scaling pitch and duration values using predefined multipliers. This strategy is interpretable and model-agnostic but does not rely on data-driven adaptation.
- **Emotion-Specific Training:** Separate FastSpeech2 models are trained from scratch on emotion-labeled subsets of the ESD corpus. Each model is optimized for a specific emotion category, allowing direct learning of emotion-specific prosodic patterns.
- **Fine-Tuning:** A FastSpeech2 model pretrained on neutral speech is fine-tuned on emotion-labeled data using lower learning rates. This transfer learning approach adapts the model to emotional prosody while retaining the base model's generalization ability.

These strategies were chosen to evaluate trade-offs in interpretability, emotional fidelity, and data efficiency.

3.5 Ethical Considerations

All datasets used in this study are publicly available and licensed for academic research purposes. None of the audio recordings contain personally identifiable information.

Participants involved in the subjective evaluation process provided informed consent. All procedures related to data collection, storage, and usage adhered to institutional ethical standards and data protection policies to ensure participant privacy and confidentiality.

4 Experimental Setup

This study aims to explore effective methods for emotional text-to-speech (TTS) synthesis based on the FastSpeech2 model. Specifically, it is hypothesized that different emotional modeling strategies can significantly enhance the emotional expressiveness and naturalness of synthesized speech. This section details the experimental settings used to validate these hypotheses, including training configurations and evaluation protocols.

4.1 Training Configuration

Three experimental setups were designed to compare alternative strategies for emotional speech synthesis using FastSpeech2.

E1: Pitch-Controlled Inference

In this setting, a base FastSpeech2 model pretrained on the LJSpeech corpus serves as the synthesis backbone. Emotional prosody is simulated during inference by manually adjusting pitch and duration values according to emotion-specific scaling factors:

- Happy: pitch $\times$ 1.3, duration $\times$ 0.8
- Sad: pitch $\times$ 0.8, duration $\times$ 1.2
- Angry: pitch $\times$ 1.2, duration $\times$ 0.9

E2: Emotion-Specific Training from Scratch

Separate FastSpeech2 models are trained from scratch for each target emotion (Happy, Sad, Angry) using subsets of the Emotional Speech Dataset (ESD). Each model is trained for 100,000 steps until convergence using randomly initialized weights.

E3: Fine-Tuning on Emotional Data

A base FastSpeech2 model pretrained on LJSpeech is fine-tuned on each emotional subset of the ESD for 16,000 steps using a reduced learning rate. This approach enables the transfer of neutral prosody to expressive targets using limited labeled data.

All training procedures were conducted using the official FastSpeech2 repository.

4.1.1 Data Preparation

To ensure consistency across all experiments, a fixed set of manually curated emotional prompts is used. The dataset is divided into 80% training, 10% development, and 10% test subsets. For each emotion, ten prompts are selected (30 in total), designed to be emotionally expressive while remaining lexically neutral. Representative examples include:

"I can't believe this is really happening to me." (happy),
"Why did this happen to me?" (sad),
"I told you this would happen, but you never listen!" (angry).

These sentences are used uniformly across conditions to control for lexical variation and focus evaluation on prosodic and emotional differences.

4.2 Evaluation Protocol

Emotional speech synthesis is evaluated using both subjective and objective measures to ensure a comprehensive assessment of naturalness and emotional expressiveness.

Subjective Evaluation: Mean Opinion Score (MOS)

A MOS test was conducted to evaluate naturalness and emotion clarity. For each emotion (Happy, Sad, Angry), three sentences were selected, resulting in nine groups. Each group contained three synthesized samples, one from each method (E1, E2, and E3), all using identical text content.
Participants were instructed to:

- Rate the naturalness of each sample on a five-point scale (1 = very unnatural, 5 = very natural)
- Select the sample that best conveyed the intended emotion

A total of 30 participants were recruited for the evaluation. All participants provided informed consent.
The evaluation was administered online using the Qualtrics platform. Each page presented a matrix of three audio samples (A, B, C) in randomized order, followed by Likert-scale naturalness ratings and a forced-choice question for emotion recognition.

Objective Evaluation

To supplement subjective analysis, acoustic prosodic features were extracted using the `openSMILE` toolkit. The following metrics were computed:

- Duration Distribution: measures consistency in utterance length across methods
- Mean F0 (Pitch): assesses alignment of synthesized pitch with emotional trends
- F0 Variance: reflects pitch dynamic range, indicative of emotional intensity

The extracted metrics are visualized using box plots, comparing synthesized outputs to the original ESD recordings and to the neutral FastSpeech2 baseline.

Through this evaluation protocol, the study aims to identify the most effective method for emotional TTS in terms of interpretability, naturalness, and data efficiency.

5 Results

This section presents the results of all experiments, including analyses of acoustic features (mean pitch, pitch variance, and utterance duration) and subjective evaluation results (naturalness and emotion recognition accuracy).

5.1 Acoustic Feature Distribution Analysis

5.1.1 Mean F0 Distribution

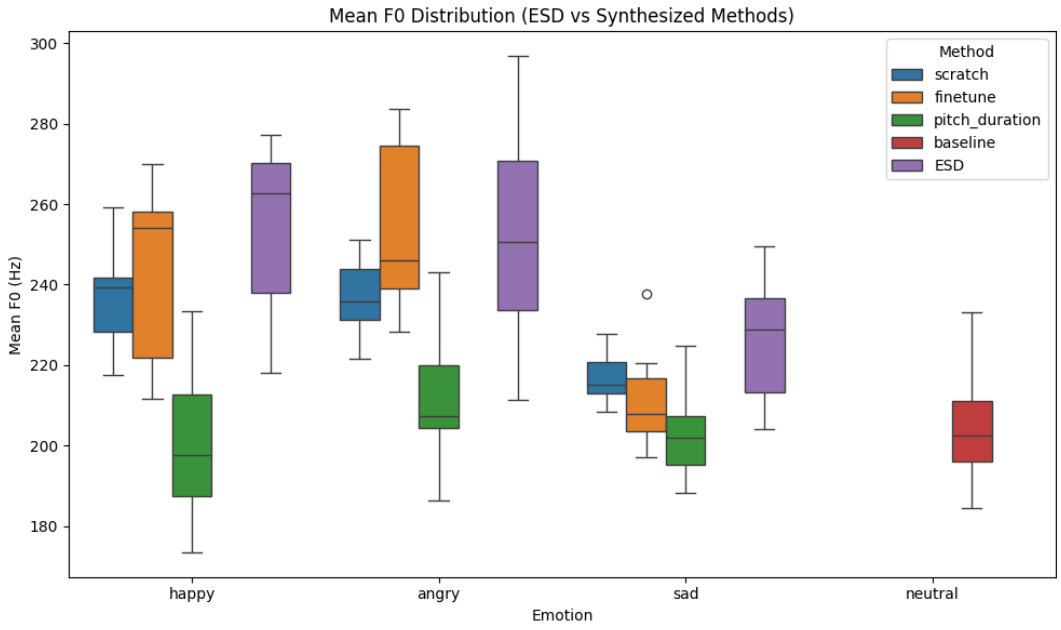

Figure 2: Mean F0 distribution across emotion categories (synthesized vs. ESD).

To evaluate emotional expressiveness, the mean fundamental frequency (F0) was analyzed across three target emotional categories: happy, angry, and sad.Synthesized speech from four methods (baseline, pitch and duration control, scratch training, and fine-tuning) was compared against natural emotional speech from the ESD corpus. Neutral speech is included only as a reference and is not discussed further.

For high-arousal emotions such as happy and angry, the fine-tuning method most accurately replicated the elevated F0 contours observed in the ESD reference, indicating strong emotional alignment. The scratch training method also captured increased pitch, but to a lesser extent. In contrast, the pitch and duration control method consistently produced lower-than-expected F0 values, failing to convey the intended emotional intensity.

In the sad category, the pitch and duration control method effectively reduced F0, closely aligning with the naturally depressed pitch profile of sad speech. However, both the scratch-trained and fine-

tuned models exhibited higher F0 than the reference, indicating less accurate emotional expression for low-arousal speech.

In summary, the fine-tuning method proved most effective in modeling emotion-specific pitch variation for high-arousal categories, while the pitch and duration control method performed better for low-arousal (sad) speech, albeit with limited expressiveness overall.

5.1.2 F0 Variance Distribution

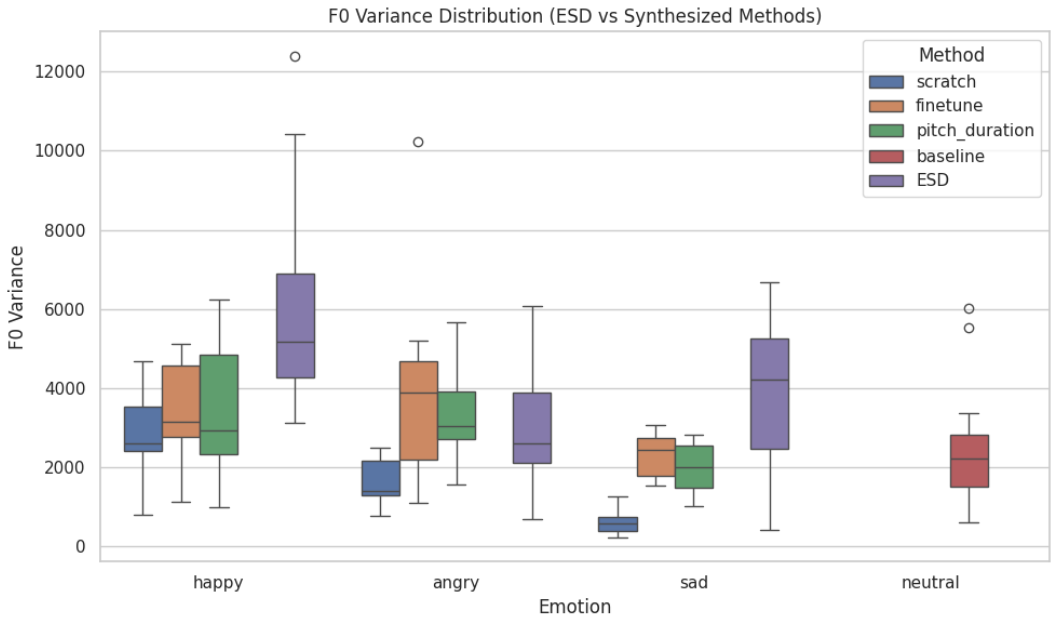

Figure 3: F0 variance distribution across emotion categories (synthesized vs. ESD).

Based on the F0 variance distributions shown in Figure 3, the fine-tuning method demonstrated the most consistent alignment with natural emotional prosody across all target emotions. For high-arousal categories such as happy and angry, the fine-tuned model approached the high variance levels observed in the ESD reference, indicating successful modeling of expressive pitch contours. In contrast, the pitch and duration control method yielded insufficient variance in these categories, resulting in notably flatter prosodic patterns.

In the sad condition, all methods appropriately exhibited reduced F0 variance in line with the ESD samples. However, the pitch and duration control method again produced the flattest profiles, suggesting limited prosodic variation even in low-arousal speech.

Overall, the fine-tuning method best captured dynamic pitch variability consistent with emotional intensity, while manual control lacked flexibility across emotional states.

5.1.3 Duration Distribution

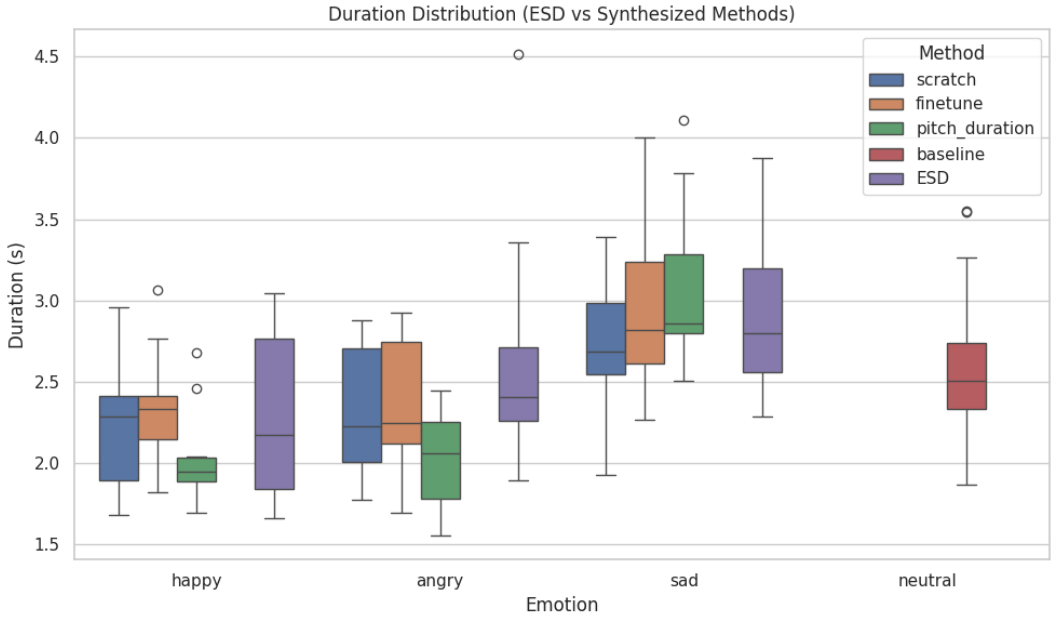

Figure 4: Utterance duration distribution across emotion categories (synthesized vs. ESD).

Utterance duration reflects speech pacing and is often modulated by emotional state. Typically, sad speech is characterized by slower and longer utterances, whereas happy and angry speech tend to be faster and shorter in duration.

As shown in Figure 4, for the sad category, the fine-tuning and scratch training methods closely approximated the longer utterance durations observed in the ESD reference, effectively modeling the slower pacing typical of low-arousal emotions. The pitch and duration control method also captured this general trend but demonstrated a more limited range of variation.

In happy and angry categories, the pitch and duration control method produced the shortest utterances among the synthesized methods, closely aligning with the ESD reference durations for high-arousal speech. However, both the fine-tuned and scratch-trained models generated noticeably longer utterances than expected, suggesting a reduced ability to compress speech rhythmically for excited emotions.

In summary, the fine-tuning method best captured slow, low-arousal pacing as seen in sad speech, while the pitch and duration control method was more effective at simulating the rapid timing associated with happy and angry emotions—though often at the expense of expressive range.

5.2 Subjective Evaluation: Naturalness and Emotion Recognition

5.2.1 Perceived Naturalness: Mean Opinion Score (MOS) Test

Figure 5 presents the average Mean Opinion Scores (MOS) grouped by synthesis method and emotion category. The fine-tuning method consistently achieved the highest naturalness ratings across all emotions, with an overall average MOS of 4.47, followed by scratch training (3.76) and pitch-duration control (3.57).

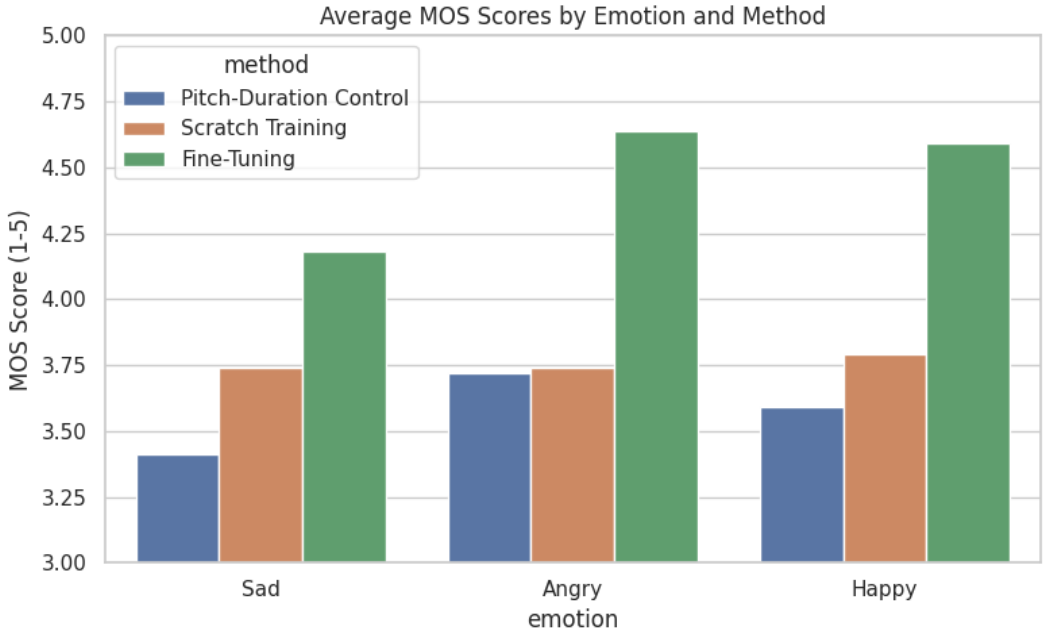

Figure 5: Average MOS Scores by Emotion and Method.

A one-way ANOVA revealed a significant main effect of synthesis method on perceived naturalness ($F = 34.15$, $p < 0.001$). Post-hoc comparisons using Tukey HSD tests (Table 1) further confirmed that the fine-tuning method significantly outperformed both the scratch training method ($p < 0.001$) and the pitch-duration control method ($p < 0.001$). However, the difference between scratch training and pitch-duration control was not statistically significant ($p = 0.2296$).

Table 1: Tukey HSD test results for synthesis methods. Statistically significant differences ($p < 0.05$) are highlighted.

Group 1	Group 2	Mean Diff	p-adj	Lower	Upper	Significant
Pitch-Duration Control	Fine-Tuning	0.8974	0.0000	0.6278	1.1670	Yes
Pitch-Duration Control	Scratch Training	0.1880	0.2296	-0.0816	0.4576	No
Fine-Tuning	Scratch Training	-0.7094	0.0000	-0.9790	-0.4398	Yes

Closer inspection of the emotional categories shows that the naturalness advantage of fine-tuning was especially prominent in the "happy" and "angry" conditions. These emotions typically demand more complex prosodic variation, such as wider pitch range, stronger intensity modulation, and faster speech rate, which fine-tuning methods appear better equipped to handle. In contrast, the difference among methods for "sad" speech was smaller, possibly due to the slower and more monotonic prosodic patterns associated with sadness, which are easier to approximate even with simpler methods.

In terms of perceptual experience, listeners may have found the fine-tuning methods smoother and more human-like, with better pitch transitions and rhythm consistency. Scratch-trained speech, while trained directly on emotional data, may have lacked sufficient exposure or variation, leading to occasional artifacts or reduced natural flow. Pitch-duration control, although interpretable and rule-based, tends to apply uniform prosodic changes that fail to capture the subtle temporal nuances and tonal variation needed for high-quality emotional synthesis.

The MOS difference of 0.89 between fine-tuning and pitch-duration control is particularly notable, representing nearly a full point on a five-point scale. This magnitude of difference reflects a substantial perceptual gain and supports the effectiveness of the fine-tuning method in modeling expressive, emotionally rich speech.

5.2.2 Emotion Recognition Accuracy

Figure 6 illustrates the results of the emotion recognition task. The fine-tuning method achieved the highest recognition accuracy (72.65%), far outperforming both the scratch training method (18.80%) and the pitch-duration control method (8.55%).

These findings underscore the fine-tuning method's superiority in capturing and reproducing salient emotional cues that are perceptible to listeners. Its ability to learn complex prosodic and spectral patterns from emotional reference data allows for more distinguishable emotional categories. While the scratch method occasionally produced expressive outputs, its lack of precise control limited emotional clarity. The rule-based pitch-duration control method, despite its simplicity, performed the worst, suggesting significant limitations in conveying fine-grained emotional nuance.

The high recognition accuracy achieved by the fine-tuning approach further validates its effectiveness not only in perceived naturalness, but also in generating emotionally distinctive speech that aligns with listener perception.

5.2.3 Correlation Between Naturalness and Emotion Recognition

To investigate the relationship between perceived naturalness and emotional clarity, Pearson correlation coefficients were computed between MOS scores and emotion recognition accuracy for each synthesis method:

- Pitch-duration control: $r = 0.996$, $p = 0.0559$

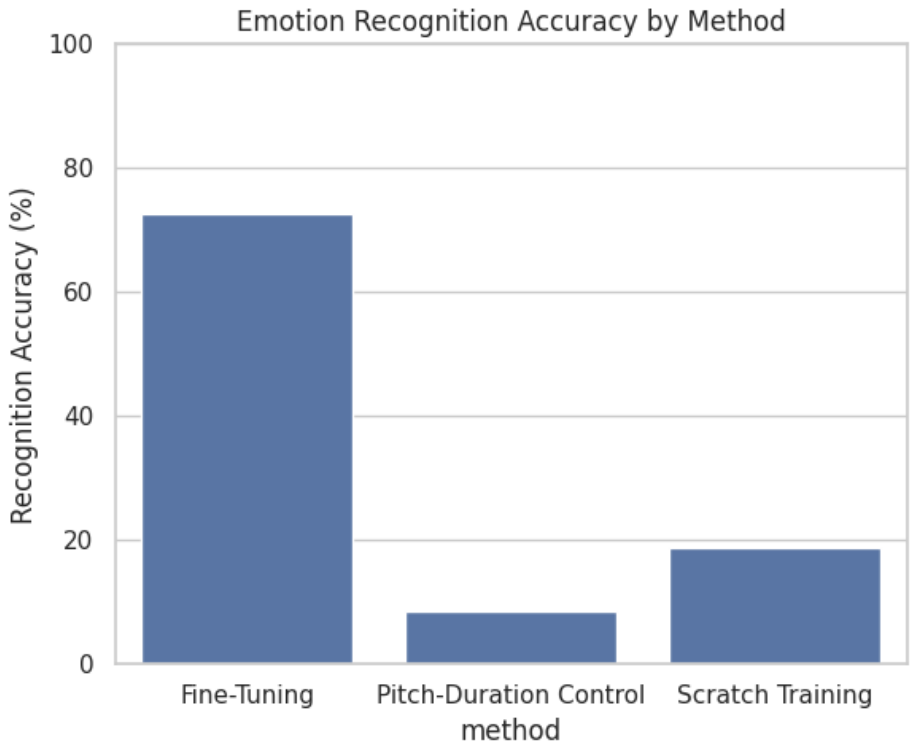

Figure 6: Emotion Recognition Accuracy by Method.

- Scratch training: $r = -0.278$, $p = 0.8208$
- Fine-tuning: $r = 0.973$, $p = 0.1491$

Although none of the correlations reached conventional statistical significance ($p > .05$), both the fine-tuning and pitch-duration control methods exhibited strong positive correlations, suggesting that more natural-sounding speech tends to be perceived as more emotionally expressive. The near-perfect correlation observed for the pitch-duration control method ($r = .996$) indicates a meaningful trend, likely limited by small sample size. In contrast, the negative correlation for the scratch training method suggests inconsistency in conveying both naturalness and emotional clarity simultaneously.

These findings support the broader hypothesis that, in expressive speech synthesis, perceived naturalness and emotional clarity are positively related. The consistent high performance of the fine-tuning method on both dimensions reinforces its advantage in emotional TTS synthesis.

6 Discussion

This section discusses the experimental findings in relation to the research hypotheses, while also addressing limitations and practical implications.

6.1 Validation of the First Hypothesis

H1 posited that both pitch-duration control and supervised emotional modeling (via from-scratch training or fine-tuning) can enhance the emotional expressiveness of synthesized speech. This hypothesis is supported by both subjective and objective evaluation results.

In terms of subjective evaluation, Mean Opinion Score (MOS) ratings showed that all three synthesis strategies were able to produce emotionally expressive speech. The fine-tuned model achieved the highest naturalness score ($M = 4.47$), followed by the from-scratch model ($M = 3.76$), and the pitch-duration controlled model ($M = 3.57$). A one-way ANOVA revealed a significant main effect of synthesis method on perceived naturalness, $F(2,N) = 34.15$, $p < .001$. Post hoc Tukey HSD tests confirmed that the fine-tuned model significantly outperformed both the from-scratch and pitch-controlled models ($p < .001$).

Objective evaluation using emotion recognition accuracy further supports this conclusion. The fine-tuned model achieved 72.65% accuracy, compared to 18.80% for the from-scratch model and 8.55% for the pitch-controlled model. These results confirm that all three methods can generate speech with perceivable emotional features, though with clear differences in effectiveness.

To summarize, the findings support H1: pitch-duration control, from-scratch emotional modeling, and fine-tuning all enhance emotional expressiveness in synthetic speech to varying degrees, with fine-tuning yielding the most consistent improvements across both subjective and objective measures.

6.2 Validation of the Second Hypothesis

H2 hypothesized that pitch-duration control would offer greater interpretability in prosodic adjustment, whereas supervised models would generate smoother and more natural prosody. The experimental results support this trade-off.

Pitch-duration control enabled manual manipulation of prosodic features such as pitch and duration. For example, higher pitch and shorter durations were assigned to high-arousal emotions like anger, while low pitch and extended durations were used for sadness. However, these modifications lacked nuanced variation and produced flat, less expressive prosody, reflected in lower naturalness scores and emotion recognition rates.

In contrast, supervised methods—particularly fine-tuning—yielded smoother pitch contours and more natural pacing, which more accurately captured the intended emotional states. The from-scratch model showed moderate expressiveness but lacked the prosodic subtlety achieved through fine-tuning.

These findings validate that while pitch control offers direct interpretability, supervised modeling—especially fine-tuning—produces higher-quality, emotionally richer speech patterns.

6.3 Validation of the Third Hypothesis

H3 proposed that fine-tuning would match or exceed the performance of from-scratch training in emotional clarity, while offering better training efficiency. The results strongly support this hypothesis.

- **MOS:** 4.47 (fine-tuned) vs. 3.76 (scratch)
- **Emotion recognition:** 72.65% vs. 18.80%
- **Training efficiency:** 16k steps vs. 100k steps

These results demonstrate the efficiency of fine-tuning with pre-trained acoustic features. Fine-tuning not only reduces the amount of required data and training steps but also yields superior results in both subjective and objective measures of emotional expressiveness.

This makes fine-tuning particularly well-suited for scenarios where emotional data is limited or computational resources are constrained. In summary, fine-tuning offers a scalable and effective solution for emotional speech synthesis, combining performance gains with practical efficiency

6.4 Limitations

Despite promising findings, several limitations merit consideration. First, the ESD data set included only 350 utterances per emotion, potentially limiting the generalizability of the results. Second, the study focused on only three emotions: happy, sad, and angry. Expanding to include a broader emotional palette (e.g., fear, surprise, disgust) would provide a more comprehensive evaluation. Third, the fine-tuning approach depends on a pre-trained neutral model, which may inherit biases from the original dataset. Fourth, the subjective evaluation involved only 30 participants. A larger and more diverse rater pool would improve the reliability and generalizability of MOS and recognition outcomes.

6.5 Summary and Implications

This study validates fine-tuning as the most effective and data-efficient strategy for synthesizing emotionally expressive speech. Although pitch control is interpretable, it lacks the fluency and expressiveness achieved by emotion embeddings. These insights can guide future TTS development, particularly in applications requiring scalable, natural, and emotionally rich voice synthesis.

7 Conclusion

This study investigated the effectiveness of different strategies to model emotional expressiveness in FastSpeech2-based text-to-speech (TTS) synthesis. The work focused on three emotion modeling strategies: manual pitch-duration control, training from scratch, and fine-tuning. In this section, we summarize the core findings and outline potential future research directions.

7.1 Summary of Contributions

A unified evaluation framework was designed to enable a controlled comparison of the three modeling strategies using consistent emotional categories and speaker data from the ESD corpus. The evaluation was subjective, using MOS scores and emotion recognition accuracy, and objective, analyzing mean F0, variance F0, and duration.

Experimental results revealed that the fine-tuned model significantly outperformed other methods in both naturalness (MOS = 4.47) and emotional clarity (72.65%). Scratch-based synthesis showed moderate effectiveness, while the pitch-duration adjustment method lagged behind.

These findings were further validated statistically. One-way analysis of variance (ANOVA) and post hoc Tukey HSD tests confirmed significant differences in perceived naturalness between methods. In addition, Pearson's correlation analyzes suggested a strong positive relationship between naturalness and emotional recognizability, especially in fine-tuned models.

7.2 Future Work

Although the current work demonstrates the value of fine-tuning for emotional speech synthesis, several directions remain open for future exploration.

First, expanding the range of emotion classes, such as fear, surprise, or disgust, could provide insights into the generalization of the model over a broader affective spectrum. Second, cross-lingual and multi-speaker adaptation remains underexplored; investigating the portability of emotional expressiveness across languages and speakers could enhance scalability.

Additionally, integrating multimodal inputs, such as semantic sentiment analysis, may lead to more expressive synthesis. Finally, optimizing emotional TTS for real-time applications, for example, in conversational agents or assistive technologies, represents a key challenge for practical deployment.

7.3 Impact and Relevance

The implications of this work extend to several real-world applications. Emotionally expressive TTS models can enhance virtual assistants, audiobook narration, accessibility solutions, and media content by making synthetic speech more human-like and engaging.

Importantly, the use of transfer learning in fine-tuning demonstrates that high-quality emotional synthesis is achievable even with limited labeled data, which is crucial for low-resource scenarios.

In conclusion, fine-tuning stands out as a data-efficient and effective approach to emotional speech synthesis. This study not only bridges the gap between interpretability and expressiveness but also lays a foundation for building emotionally aware speech systems that align more closely with human communicative expectations.

References

Bulut, M., Busso, C., Yildirim, S., Kazemzadeh, A., Lee, C. M., Lee, S., & Narayanan, S. (2005). Investigating the role of phoneme-level modifications in emotional speech resynthesis. In *Interspeech 2005* (pp. 801–804). doi: 10.21437/Interspeech.2005-378

Burkhardt, F., & Campbell, N. (2015). Emotional speech synthesis. In R. Calvo et al. (Eds.), *The oxford handbook of affective computing.* Oxford University Press. doi: 10.1093/oxfordhb/9780199942237.013.038

Cen, L., Dong, M., & Chan, P. Y. (2011). Data pre-processing in emotional speech synthesis by emotion recognition. In *Proceedings of the international conference on affective computing and intelligent interaction (acii).* Memphis, TN, USA: IEEE.

Cornille, T., Wang, F., & Bekker, J. (2022). Interactive multi-level prosody control for expressive speech synthesis. In *Icassp 2022 - 2022 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 8312-8316). doi: 10.1109/ICASSP43922.2022.9746654

Diatlova, D., & Shutov, V. (2023). *Emospeech: Guiding fastspeech2 towards emotional text to speech.* Retrieved from `https://arxiv.org/abs/2307.00024`

Fahad, M. S., Singh, S., Gupta, S., Deepak, A., & Abhinav, A. (2019). Synthesis of emotional speech by prosody modification of vowel segments of neutral speech.

Govind, D., & Prasanna, S. R. (2012). Dynamic prosody modification using zero frequency filtered signal. *International Journal of Speech Technology*, *16*, 41–54. doi: 10.1007/s10772-012-9141-0

He, K., Sun, C., Zhu, R., & Zhao, L. (2022). Multi-speaker emotional speech synthesis with limited datasets: Two-stage non-parallel training strategy. In *2022 7th international conference on intelligent computing and signal processing (icsp)* (p. 545-548). doi: 10.1109/ICSP54964.2022.9778768

Hoult, C. (2004). *Emotion in speech synthesis.* (Unpublished manuscript or internal report)

Inoue, S., Zhou, K., Wang, S., & Li, H. (2024). Fine-grained quantitative emotion editing for speech generation. In *2024 asia pacific signal and information processing association annual summit and conference (apsipa asc)* (p. 1-6). doi: 10.1109/APSIPAASC63619.2025.10848721

Kalita, J., & Deb, N. (2017, April). Emotional text to speech synthesis: A review. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, *6*, 428–430. doi: 10.17148/IJARCCE.2017.6482

Kim, J.-K., soo Hahn, H., Yoo, U.-J., & Bae, M.-J. (2008). On a pitch duration technique for prosody control. *World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering*, *2*, 1858–1861.

Koike, K., Suzuki, H., & Saito, H. (1998). Prosodic parameters in emotional speech. In *5th international conference on spoken language processing (icslp 1998)* (p. paper 0996). doi: 10.21437/ICSLP.1998-136

Kolekar, S. S., Richter, D. J., Bappi, M. I., & Kim, K. (2024). Advancing ai voice synthesis: Integrating emotional expression in multi-speaker voice generation. In *2024 international conference on artificial intelligence in information and communication (icaiic)* (p. 193-198). doi: 10.1109/ICAIIC60209.2024.10463204

Lee, Y., & Kim, T. (2019). Robust and fine-grained prosody control of end-to-end speech synthesis. In *Icassp 2019 - ieee international conference on acoustics, speech and signal processing* (pp. 5911–5915). doi: 10.1109/ICASSP.2019.8683501

Lei, Y., Yang, S., Wang, X., & Xie, L. (2022). Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 853-864. doi: 10.1109/TASLP.2022.3145293

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017* (pp. 498–502). doi: 10.21437/Interspeech.2017-1386

Nithin, S. K., & Prakash, J. (2022). Emotional speech synthesis using end-to-end neural tts models. In *2022 18th international computer engineering conference (icenco)* (Vol. 1, p. 1-7). doi: 10.1109/ICENCO55801.2022.10032463

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). *Fastspeech 2: Fast and high-quality end-to-end text to speech.* Retrieved from https://arxiv.org/abs/2006.04558

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). *Fastspeech: Fast, robust and controllable text to speech.* Retrieved from https://arxiv.org/abs/1905.09263

Schröder, M. (2001). Emotional speech synthesis: A review. In *7th european conference on speech communication and technology (eurospeech 2001)* (pp. 561–564). doi: 10.21437/Eurospeech.2001-150

Thi, N.-A. N., Thang Ta, B., Le, N. M., & Hai Do, V. (2023). An automatic pipeline for building emotional speech dataset. In *2023 asia pacific signal and information processing association annual summit and conference (apsipa asc)* (p. 1030-1035). doi: 10.1109/APSIPAASC58517.2023.10317420

Tits, N., Haddad, K. E., & Dutoit, T. (2019). Emotional speech datasets for english speech synthesis purpose: A review. *Intelligent Systems with Applications*.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... Saurous, R. A. (2017). *Tacotron: Towards end-to-end speech synthesis.* Retrieved from https://arxiv.org/abs/1703.10135

Wong, K. Y., & Chung, K. F. (2024). Pico-vits: Leveraging pitch contours for fine-grained emotional speech synthesis. In *Proceedings of the international conference on text, speech and dialogue.*

Zhou, P., Chong, D., Wang, H., & Zeng, Q. (2022). Calibrate and refine! a novel and agile framework for asr error robust intent detection. In *Interspeech 2022* (pp. 1096–1100). doi: 10.21437/Interspeech.2022-786

Łańcucki, A. (2021). Fastpitch: Parallel text-to-speech with pitch prediction. In *Icassp 2021 - ieee international conference on acoustics, speech and signal processing* (pp. 6588–6592). doi: 10.1109/ICASSP39728.2021.9413889

A Questionnaire Survey

A.1 Questionnaire Design

To evaluate the perceptual quality of synthesized emotional speech, a structured listening test was conducted using an online questionnaire. The test consisted of 40 items, divided into two types of evaluation:

- **Naturalness Rating:** Listeners rated the naturalness of each synthesized audio sample on a five-point Likert scale (1 = very unnatural, 5 = very natural).
- **Emotion Recognition:** For each target emotion (sad, angry, happy), the participants selected the version that most clearly conveyed the intended emotion, based on prosodic cues such as pitch, rhythm, and duration.

Each item involved a randomized triplet of audio samples labeled A, B, or C, corresponding to the following synthesis methods:

- adjust_pitch_and_duration
- scratch
- finetune

The ordering of A/B/C options was randomized across questions to mitigate positional bias.

A.2 Participant Instructions

Participants were presented with an overview of the study objectives, procedures, and informed consent form at the beginning of the questionnaire. The test was implemented using the Qualtrics platform and took approximately 10–15 minutes to complete. Participants were instructed to complete the task in a quiet environment, preferably using headphones to ensure optimal listening conditions. A screenshot of the welcome screen is shown in Figure 7.

A.3 Question Format

Each evaluation item consisted of a set of three audio clips and two corresponding questions: a naturalness rating and an emotion recognition task. The interface presented all three versions simultaneously, allowing direct comparison. A screenshot of the typical interface layout is shown in Figure 8.

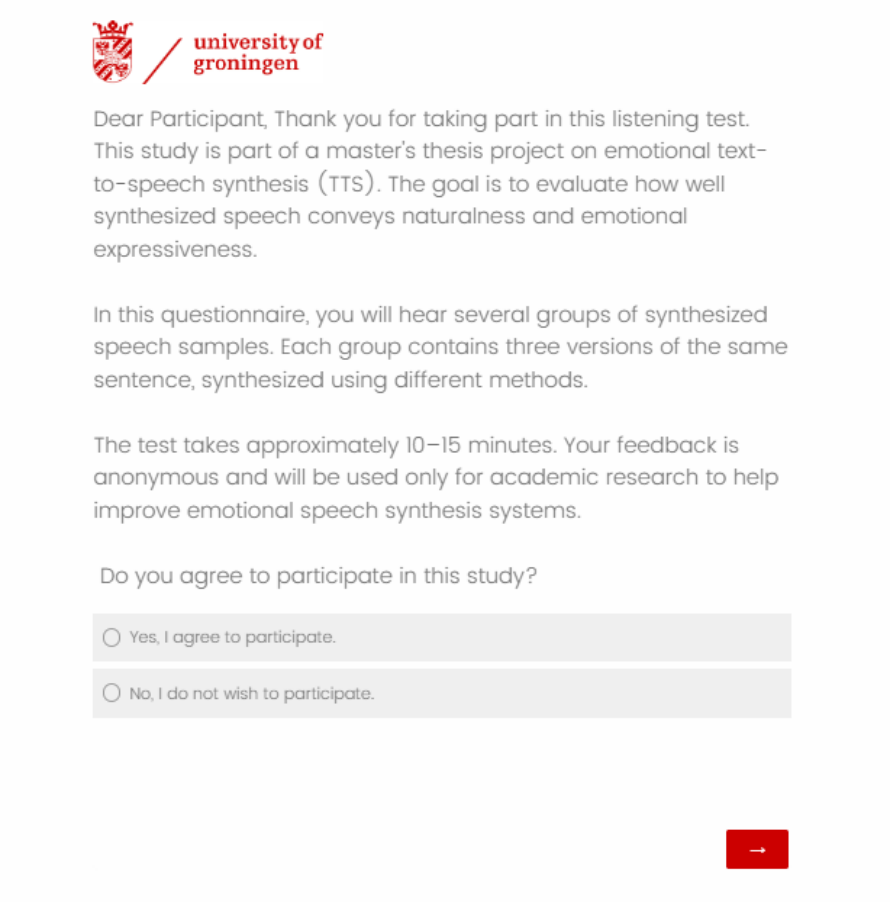

Figure 7: Participant welcome screen with study description and consent option.

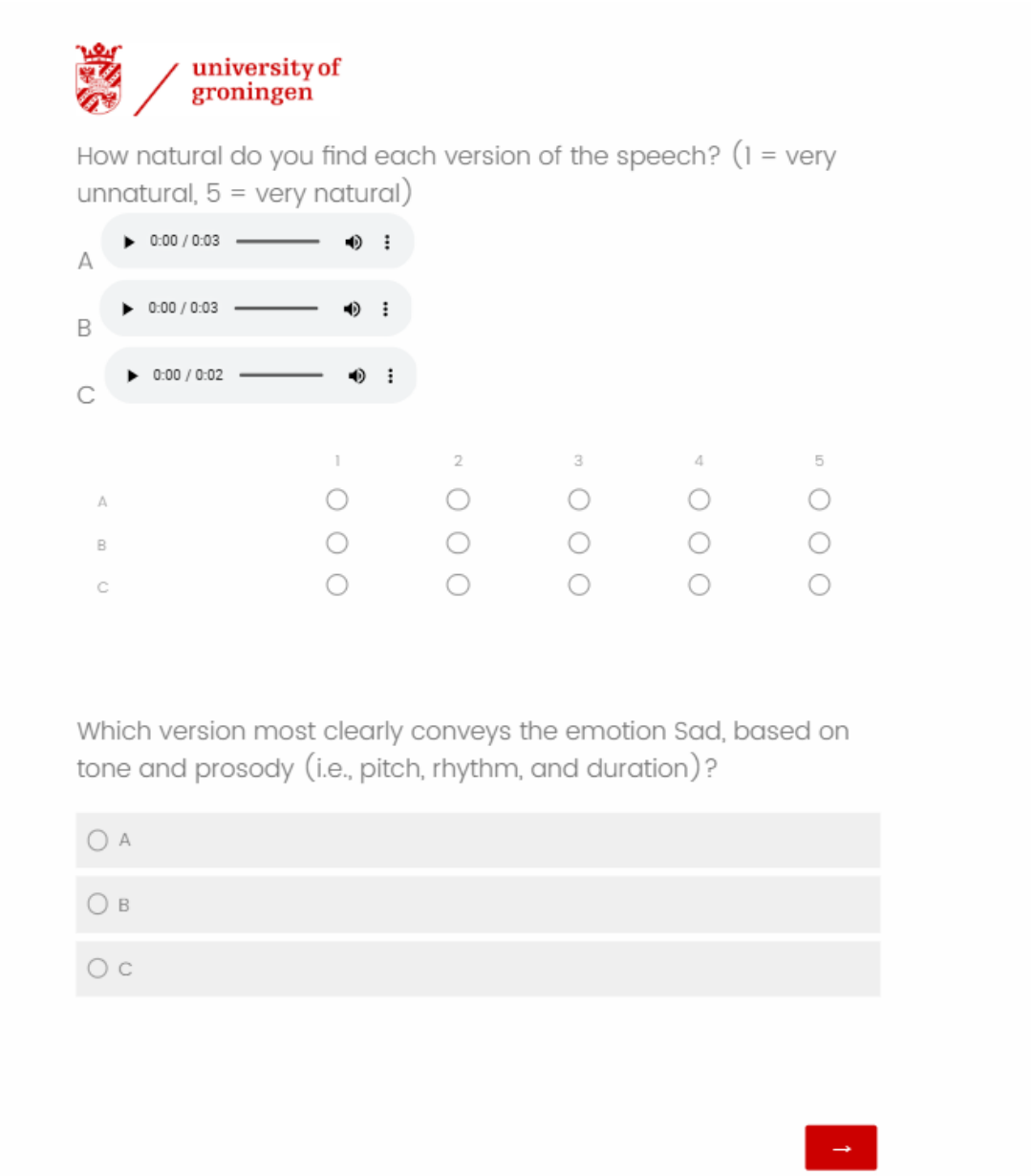

Figure 8: Screenshot of the perceptual evaluation interface, showing naturalness ratings and emotion recognition selection.

A.4 Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. During the preparation of this thesis, I used OpenAI ChatGPT-4 for the following purposes: Summarizing background literature for preliminary review, specifically in Section 2.4, "Emotional Speech Datasets for English Speech Synthesis Purpose: A Review"; Assisting with formatting in-text citations and references according to the APA style. All content was subsequently reviewed, verified, and substantially modified by me to ensure accuracy, relevance, and alignment with academic standards.